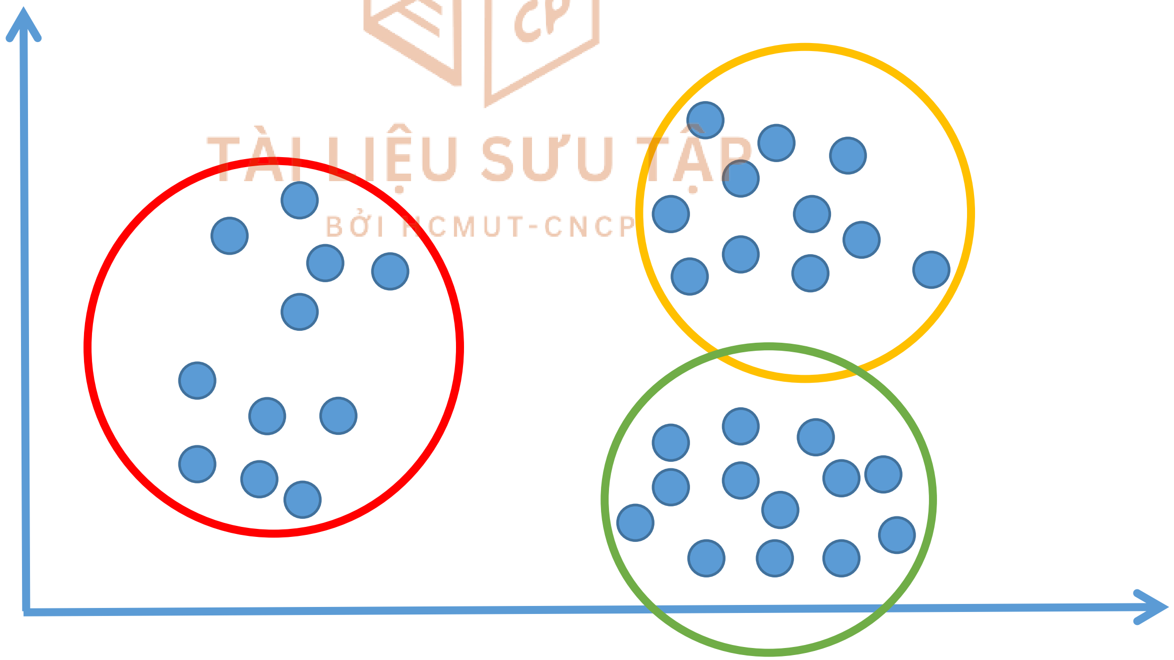
Artificial Intelligence

k - means Clustering

Pham Viet Cuong
Dept. Control Engineering & Automation, FEEE
Ho Chi Minh City University of Technology

k - Means Clustering

- Basic idea: group together similar instances
  - High intra-cluster similarity
  - Low inter-cluster similarity



k - Means Clustering

- Example:
  - Document clustering
    - Web search engine often return thousands of pages --> Difficult for user
    - Clustering can be used to group retrieved documents into categories
  - Customer segmentation
  - Recommendation engines
  - Image compression

k - Means Clustering

✓ Supervised or unsupervised?

Supervised Classification	Unsupervised Clustering
• known number of classes	• unknown number of classes
• based on a training set	• no prior knowledge
• used to classify future observations	• used to understand (explore) data

✓ Requires data, but no labels

✓ Useful when don’t know what we’re looking for

k - Means Clustering

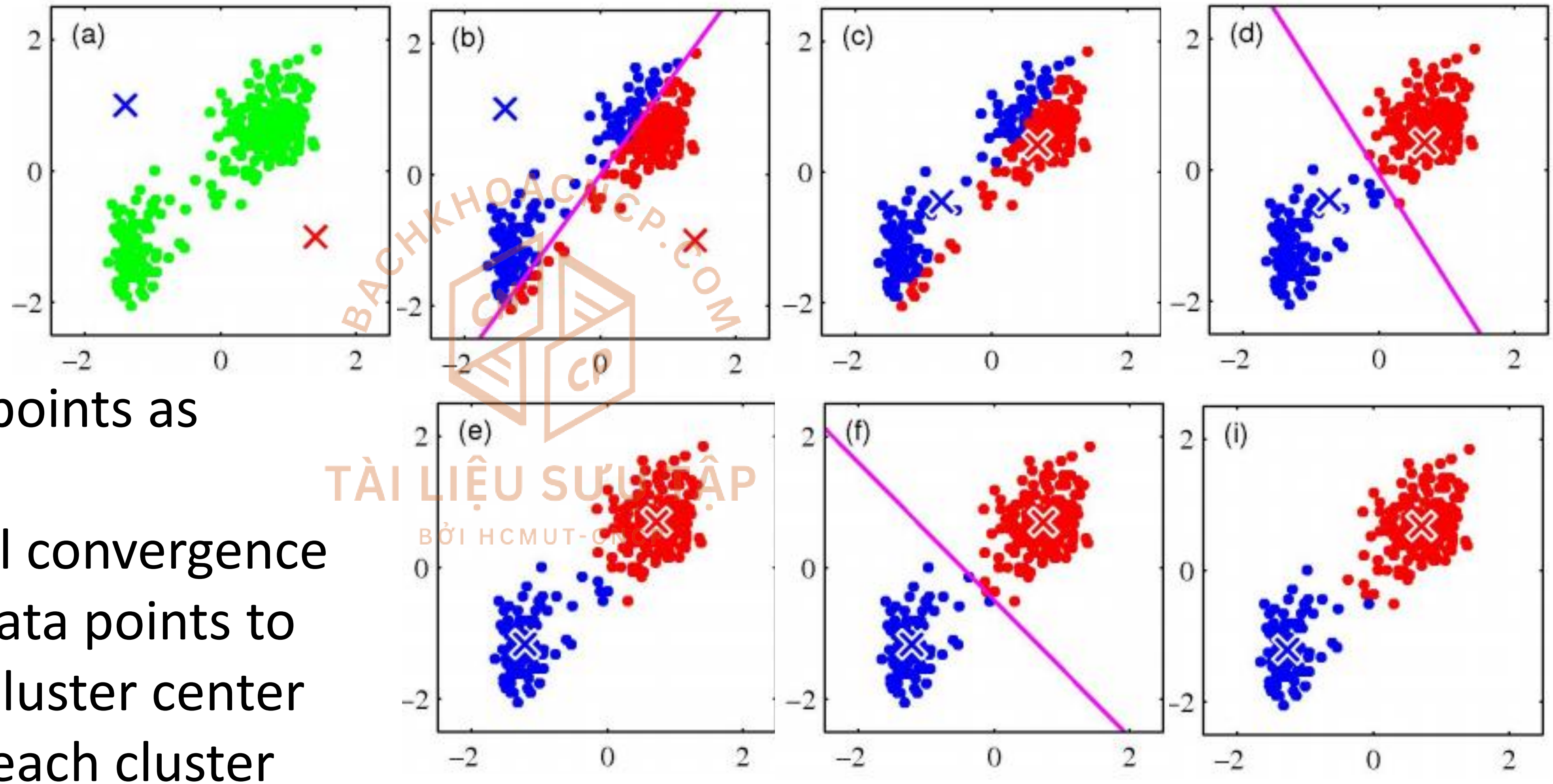
- Requirements
  - An integer k
  - A set of training data (without labels)
  - A metric to measure similarity

- Algorithm
  - Pick k random points as cluster centers
  - Repeat until convergence
    - Assign data points to closest cluster center
    - Update each cluster center to be the mean of its assigned points

Convergence: No points' assignments change

k - Means Clustering

✓ Example 1



✓ Pick k random points as cluster centers

❖ Repeat until convergence

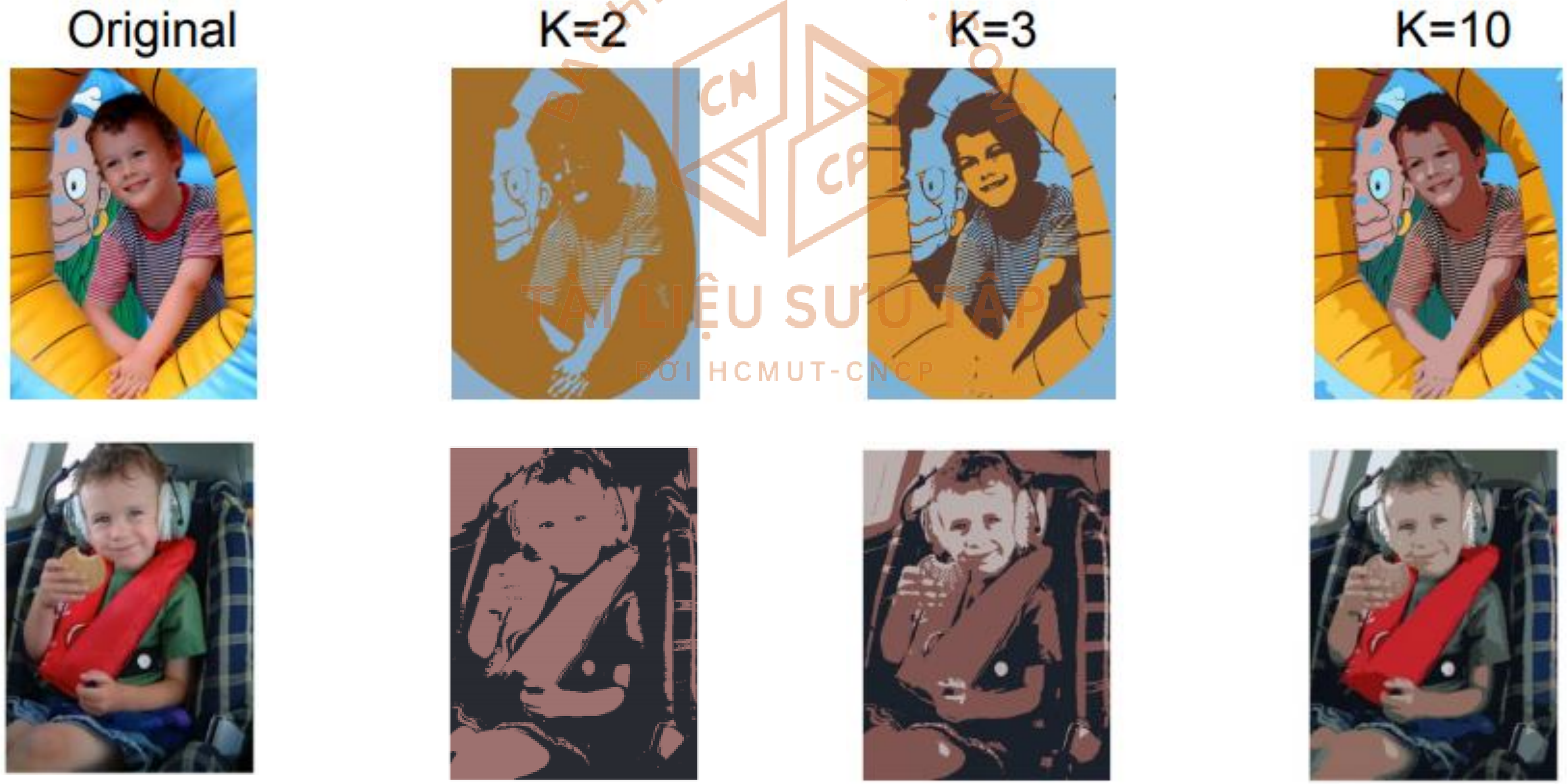
- Assign data points to closest cluster center
- Update each cluster center to be the mean of its assigned points

k - Means Clustering

- Example 2
- Example 3

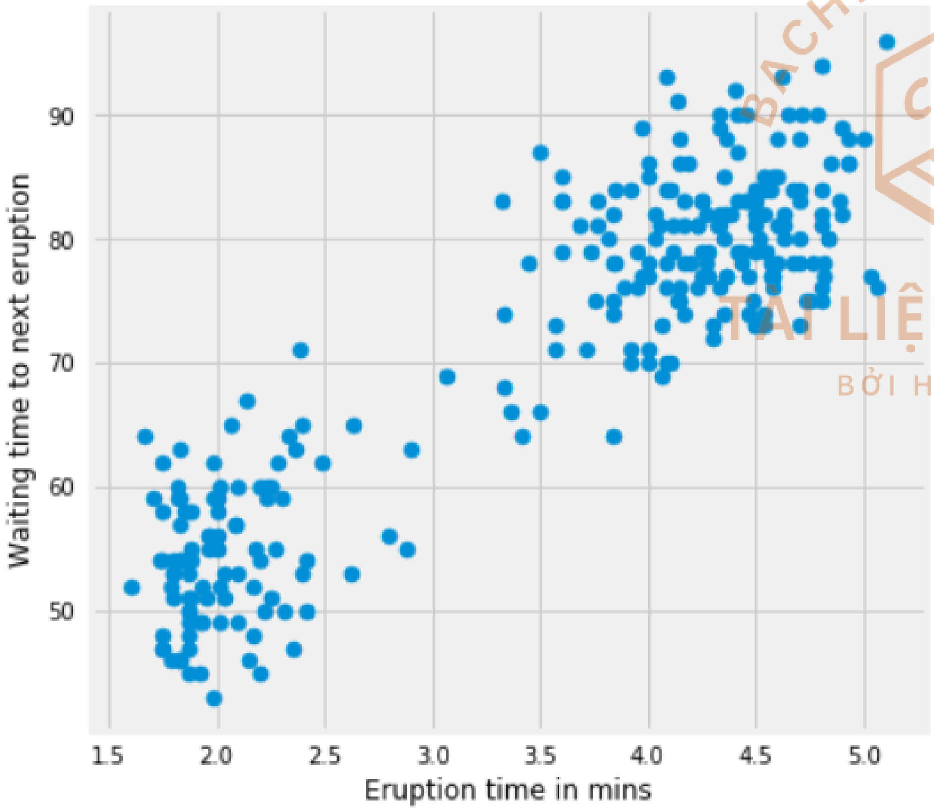
k - Means Clustering

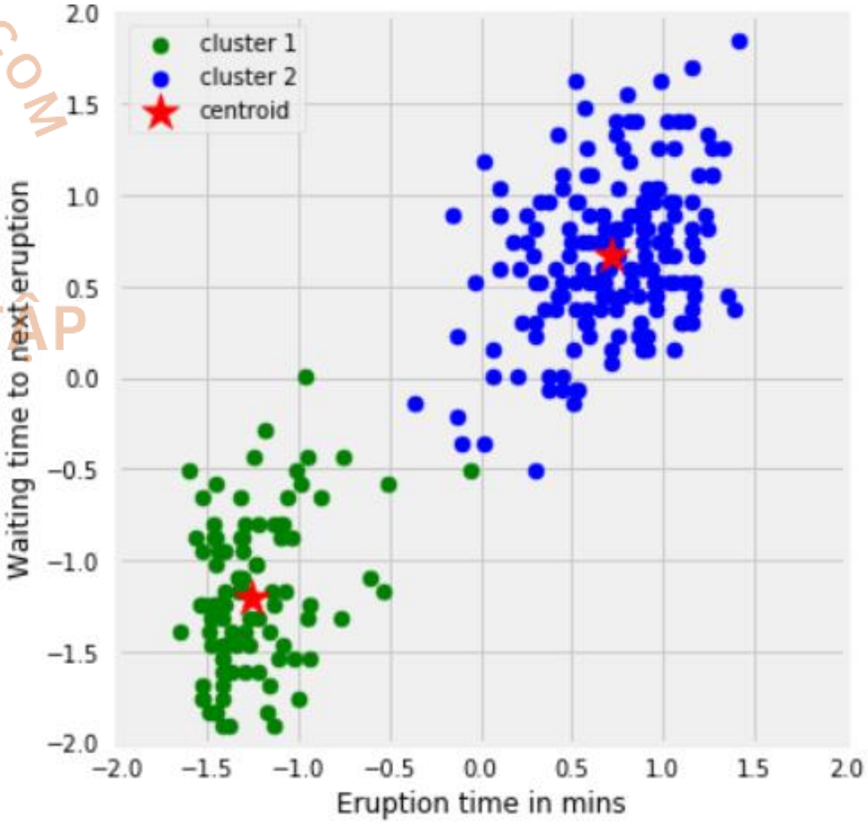
- Example: Image segmentation
  - Segmentation: partition an image into regions each of which has reasonably homogenous visual appearance



k - Means Clustering

- Example: Geyser eruptions
  - Eruption time (mins)
  - Waiting time to next eruption (mins)





k - Means Clustering

- Example: Image compression
  - Original image: 396*396*24 = 3,763,584 bits
  - Compressed image: 30*24 + 396*396*5 = 784,800 bits

k - Means Clustering

- Properties
  - Guaranteed to converge in a finite number of iterations
  - Running time per iteration
    - Assign data points to closest cluster center

      O(kN)
    - Update the cluster center to be the mean of its assigned points

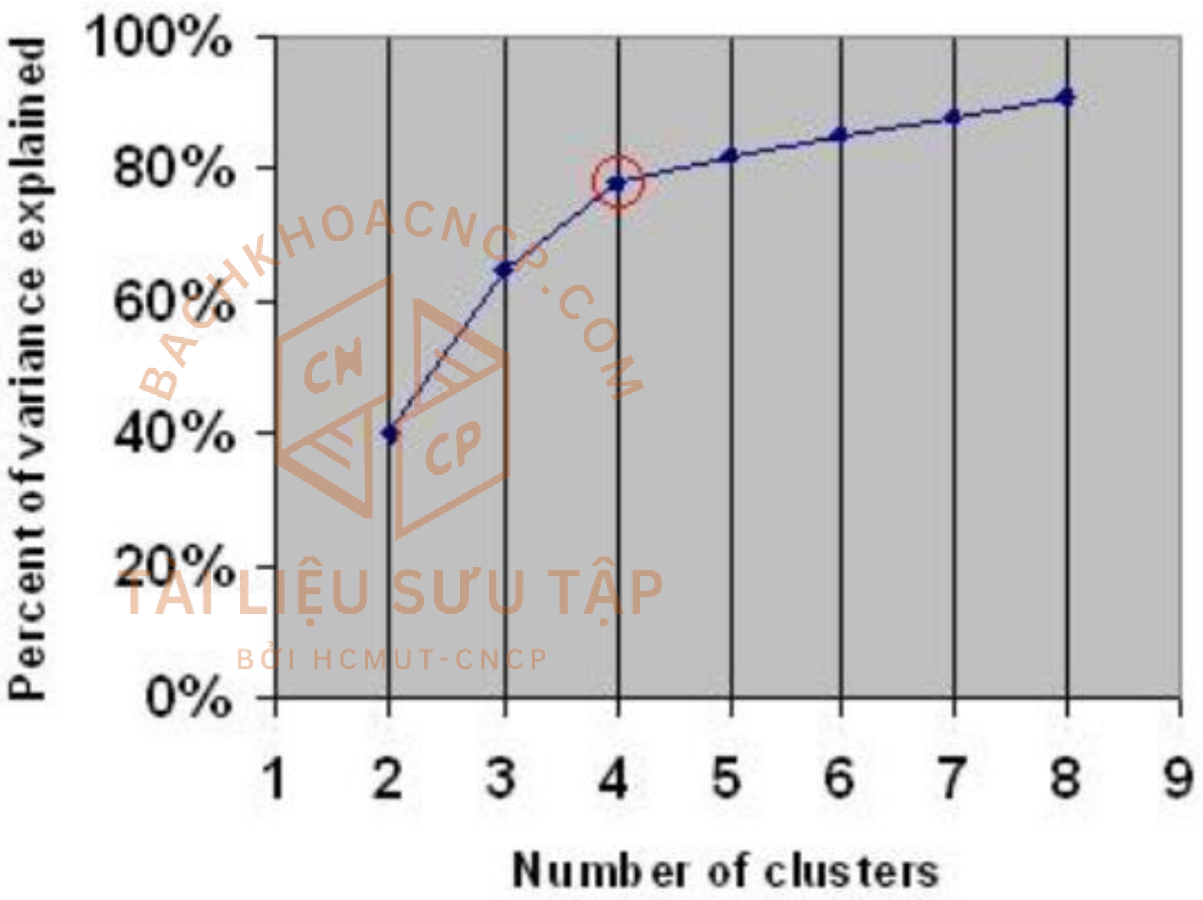
      O(N)

k - Means Clustering

- How to measure similarity?
  - Similarity is subjective
  - Depends on data, cases, users, etc.
  - Not always straightforward which metrics work well
  - “Trial and error” can be used
  - Examples of similarity measures: Euclidean, Mahattan, cosine distance

k - Means Clustering

✓ How to choose k?
  ❖ Elbow method



Percentage of variance explained is the ratio of the between-group variance to the total variance

k - Means Clustering

✓ How to initialize centroids?

❖ K-means ++

The intuition behind this approach is that spreading out the *k* initial cluster centers is a good thing: the first cluster center is chosen uniformly at random from the data points that are being clustered, after which each subsequent cluster center is chosen from the *remaining* data points with probability proportional to its squared distance from the point's closest existing cluster center.

The exact algorithm is as follows:

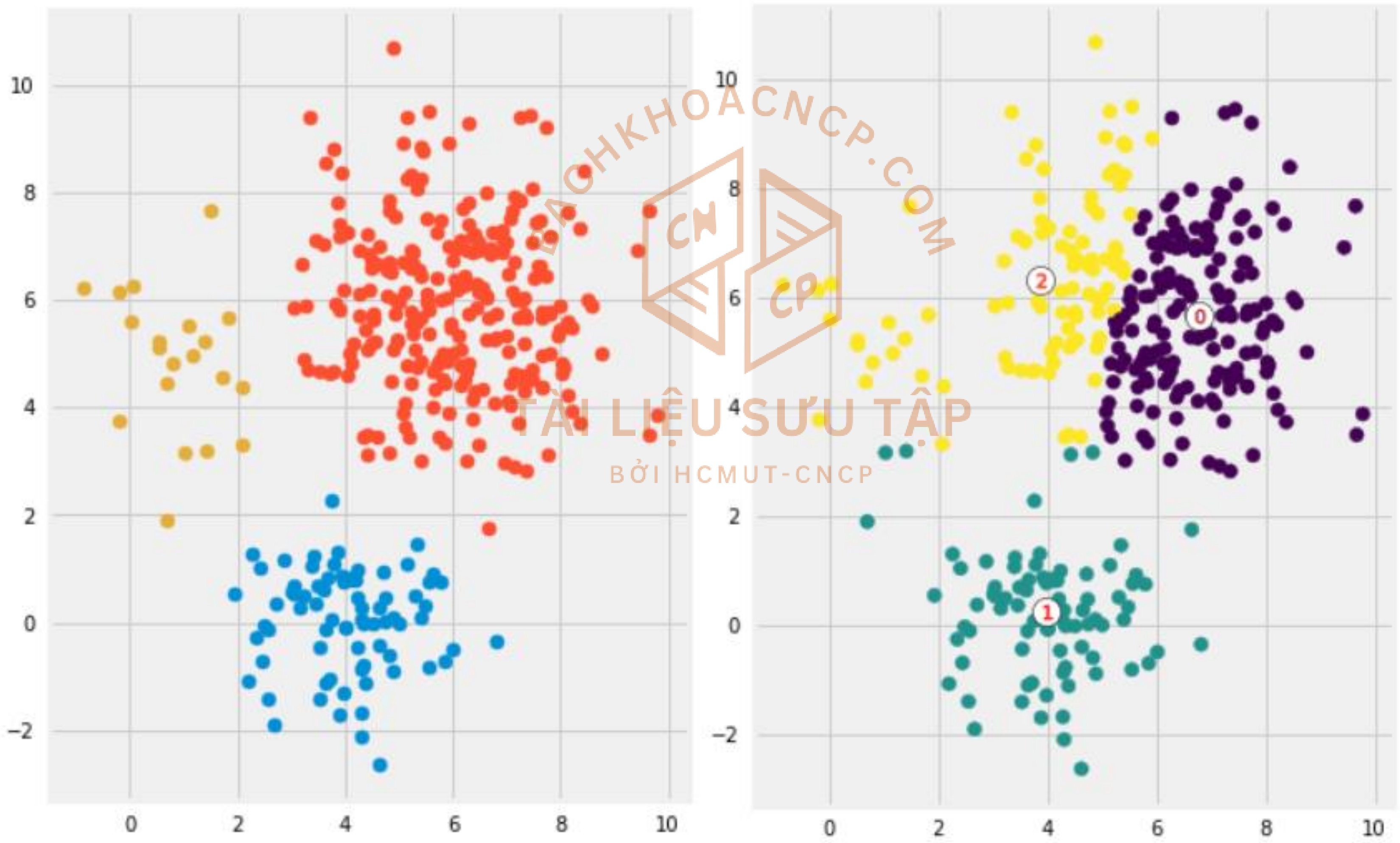
1. Choose one center uniformly at random among the data points.
2. For each data point x not chosen yet, compute D(x), the distance between x and the nearest center that has already been chosen.
3. Choose one new data point at random as a new center, using a weighted probability distribution where a point x is chosen with probability proportional to D(x)^2.
4. Repeat Steps 2 and 3 until k centers have been chosen.
5. Now that the initial centers have been chosen, proceed using standard *k*-means clustering.

k - Means Clustering

- k-means clustering: heuristic
  - Requires initial means
  - Does matter what you pick

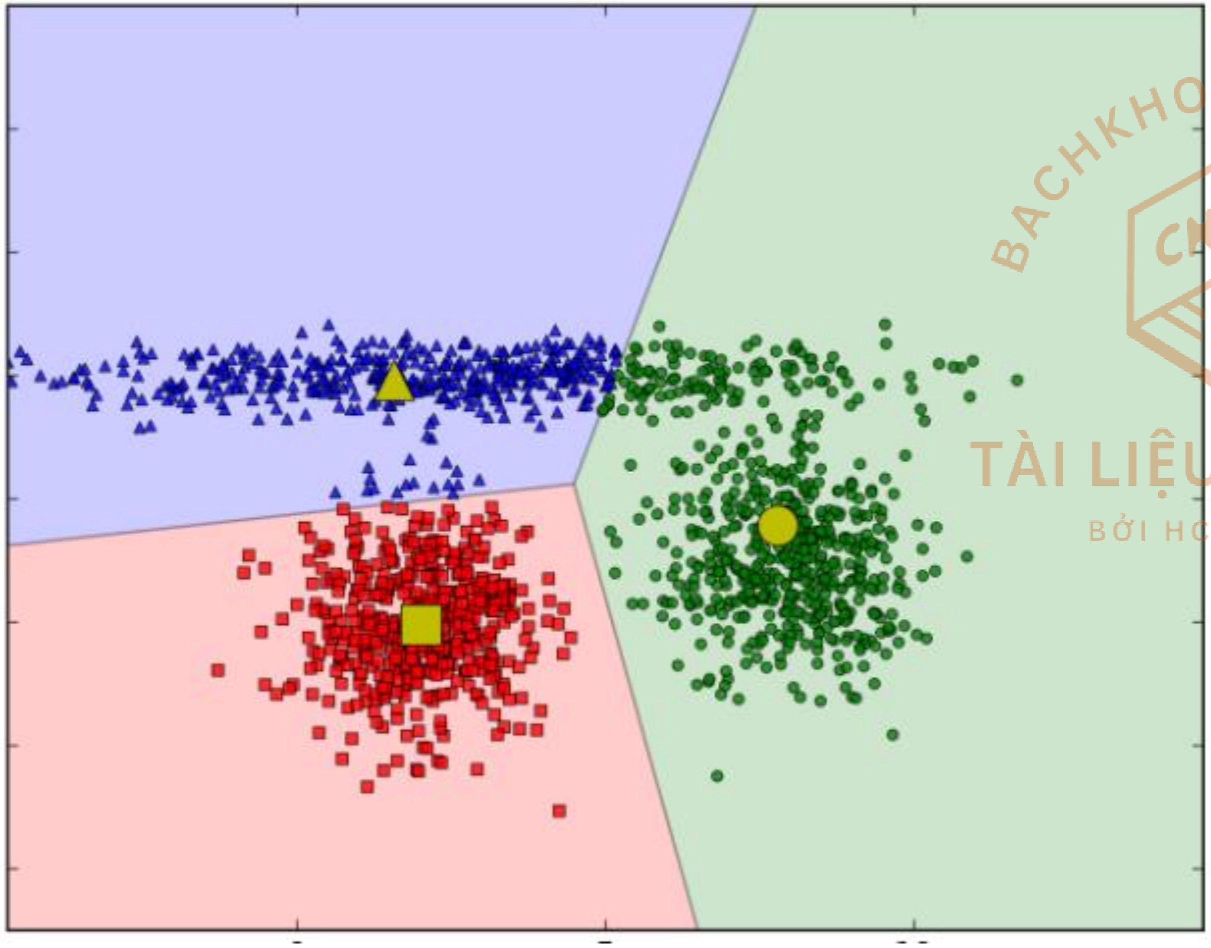
k - Means Clustering

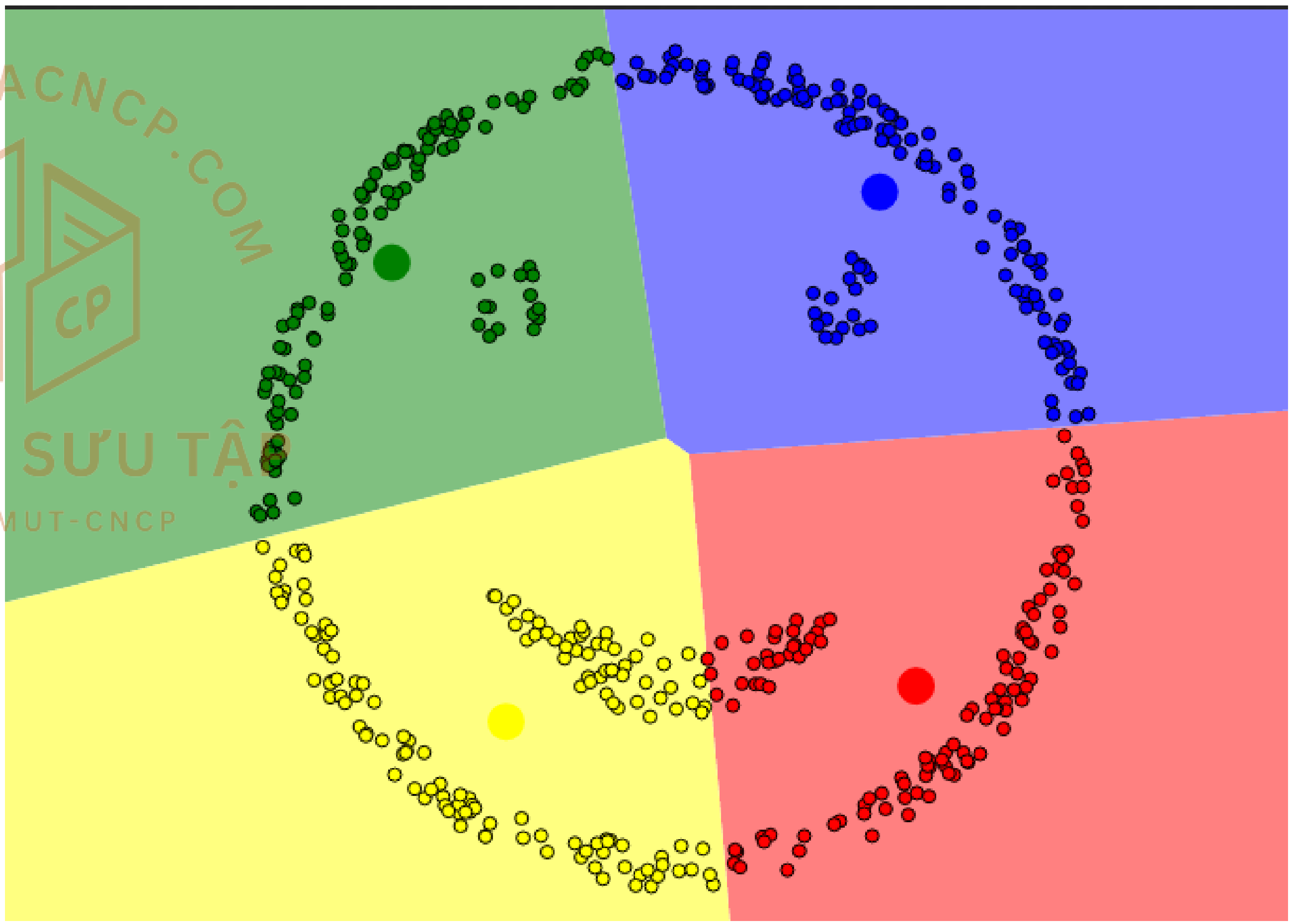
✓ Drawbacks



k - Means Clustering

✓ Drawbacks





Silhouette (clustering)

From Wikipedia, the free encyclopedia

Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object has been classified.[1]

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

The silhouette can be calculated with any distance metric, such as the Euclidean distance or the Manhattan distance.

k - Means Clustering

For data point $i \in C_i$ (data point i in the cluster C_i), let

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

For each data point $i \in C_i$, we now define

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j)$$

We now define a *silhouette* (value) of one data point i

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_i| > 1$$

and

$$s(i) = 0, \text{ if } |C_i| = 1$$

Which can be also written as:

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$

From the above definition it is clear that

$$-1 \le s(i) \le 1$$

k - Means Clustering

- Sources:
  - http://people.csail.mit.edu/dsontag/courses/ml12/slides/lecture14.pdf
  - https://www.slideshare.net/annafensel/kmeans-clustering-122651195
  - https://en.wikipedia.org/wiki/Elbow_method_(clustering)
  - https://www2.stat.duke.edu/courses/Fall02/sta290/datasets/geyser